DOCUMENT RESUME

ED 345 498 FL 019 628

AUTHOR McNamara, T. F.; Adams, R. J.
TITLE Exploring Rater Behaviour with Rasch Techniques.
PUB DATE Mar 91
NOTE 29p.; Paper presented at the Annual Language Testing Research Colloquium (Princeton, NJ, March 21-23, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; *English (Second Language); Foreign Countries; *Interrater Reliability; *Language Proficiency; *Language Tests; *Measurement Techniques; Models; Rating Scales; Second Language Learning; Testing; Test Reliability
IDENTIFIERS *Rasch Model

ABSTRACT

A preliminary study is reported of the use of new multifaceted Rasch measurement mechanisms for investigating rater characteristics in language testing. Ratings from four judges of scripts from 50 candidates taking the International English Language Testing System test, a test of English for Academic Purposes, are analyzed. The analysis illustrates how multifaceted Rasch measurement can be used to examine inter-rater consistency, differences in rater harshness, available grades on the rating scale, and the effect that between-rater variation has on the measurement of individual candidates. Although the main focus of the paper is on modeling and estimating rater variation, Rasch modeling also has the potential for practical applications controlling the effects of the variation it describes. One such application is considered: the use of the model to explore the relationship between varying amounts of multiple marking and the resulting ability estimates of candidates, to see if it may be possible to reduce the amount of multiple marking required to produce stable and reliable estimates of ability. Contains 18 references. (LB)

EXPLORING RATER BEHAVIOUR WITH RASCH TECHNIQUES[1]

T.F. McNAMARA (University of Melbourne)
and
R.J. ADAMS (Australian Council for Educational Research)

1 Introduction

One of the consequences of the widespread acceptance of communicative approaches to language testing has been an increased use of subjective assessments to measure performance on realistic written and spoken tasks. This has led in turn to a corresponding need to establish the reliability and validity of such assessments. A difficulty arises in these situations because raters contribute an additional source of variation to the measurements (additional, that is, to the variation associated with test items). This rater variation can be considerable, and cannot be ignored. The extent and nature of rater variation is currently *estimated* in terms of inter- or intra-rater reliabilities, or, less commonly, using generalizability coefficients (Bachman, 1990) or other ANOVA-based procedures (Woods and Krzanowski, 1984; Woods, Fletcher and Hughes, 1986). Current procedures for *controlling* this additional source of variation typically include multiple rating of scripts in the case of writing assessments or tapes or even live performances in the case of the speaking skill. Such procedures are inevitably expensive.

Recent developments in multi-faceted Rasch measurement (Linacre, 1989) provide improved mechanisms for investigating rater characteristics. In this paper a preliminary study of the use of these new techniques in the analysis of test data is presented. We describe the analysis of ratings from four judges of scripts from 50 candidates taking the International English Language Testing System (IELTS) test, a test of English for Academic Purposes. The analysis illustrates how multi-faceted Rasch measurement can be used to examine inter-rater consistency, differences in rater harshness, differences in the manner in which raters use the available grades on the rating scale and the effect that between rater variation has on the measurement of individual candidates.

Although the main focus of the paper is on modelling and estimating rater variation, Rasch modelling also has the potential for practical applications in terms of controlling for the effects of the variation it describes. In the latter part of the paper, one such application is considered: the use of the model to explore the relationship between varying amounts of multiple marking and resulting ability estimates of candidates, to see if it may be possible to reduce the amount of multiple marking that may be required to produce stable and reliable estimates of ability.

2 Rasch models and facets

Rasch Item Response Theory comprises a family of models (Masters and Wright, 1984). The basic model (Wright and Stone, 1979) handles dichotomously scored items. Rating Scale analysis (Andrich, 1978a, Andrich, 1978b), an extension of the basic model, can handle data from Likert-type scales. The Partial Credit Model (Wright and Masters, 1982; Masters, 1982), enables analysis of items in which a range of marks may be awarded to a response, depending on its quality. With *dichotomous* data, estimates are available of the likelihood of a candidate of a certain ability getting an item of given difficulty right or wrong. With *Rating Scale analysis*, this is extended to the likelihood of a given candidate achieving a certain score on a scale for an item of given difficulty. *Partial Credit analysis* also provides information on rating scale thresholds for individual items, that is, how difficult it is for a candidate of given ability to move from one score point to another on that item. This analysis in other words allows one to explore the structure of the rating scale: that is, it makes fewer assumptions about how the rating scale is being interpreted. Thus, Rasch models allow one to make probabilistic statements about item difficulty, candidate ability and rating scale thresholds. Such statements are expressed in terms of units called logits, the logarithm of the odds of a certain outcome.

Recently, Linacre (1989:48-49) has described the extension of the Rating Scale and Partial Credit models to include judge characteristics:

> *With the inclusion of the judges in the measurement process, it is useful to define simultaneously not only the ability of the examinee, and the difficulty of the test item, but also the severity of the judge. This is accomplished by expanding the rating scale model to include parameters describing each judge's method of applying the rating scale. This introduces the additional facet of judges into the previous framework of examinees and items. The traditional paper-and-pencil test is a two-faceted test, and the intermediation of a judge makes the testing situation into a three-faceted one.*

Thus according to Linacre (1989: 51)

> *Judge severity, item difficulty, examinee ability and rating scale thresholds can be expressed as real numbers on a common interval scale. They combine additively to provide an expectation of the logarithm of the odds of a judge awarding a rating in one category to an examinee's performance on an item, as compared to that same judge awarding the rating in the next lower category.*

Linacre has developed the necessary software, FACETS (Linacre, 1990). This software was used in the analysis reported below.

Expectations of particular outcomes can be expressed graphically by means of Item Characteristic Curves. These allow inspection of the likelihood of a candidate of a particular ability achieving a particular score on an item of given difficulty, with a judge of given harshness. (In this paper, as the judging task involves awarding separate marks to written scripts on a number of aspects or dimensions of performance such as *Task fulfilment*, *Coherence and cohesion* and *Sentence structure*, reference will be made to *performance dimensions* rather than to 'items'.)

3 Rater characteristics

There is an ambiguity in Linacre's use of the term 'judge severity'. On the one hand, it can be interpreted as an overall characteristic of the judge. On the other hand, particular performance dimensions may elicit patterns of harshness that may be different from those elicited by other such dimensions; in other words, there may be an *interaction* between judges and performance dimensions. Additionally, there may be an interaction between judges and the rating scale thresholds, so that the thresholds may be interpreted differently by different judges. There may be a three-way interaction between judges, performance dimensions and thresholds. (The distinction being drawn here is rather like that between main effects and interaction effects in ANOVA.) Rather than 'judge severity', then, we prefer to use the term *rater characteristics*, a number of which may be analysed using the new model.

Assume for example that ratings are being given on a number of separate aspects of performance on a language task. Ratings are being assigned in each performance dimension on let us say a five-point scale (cf Figure 1).

Figure 1 Rating scale

Performance dimension 1 0 1 2 3 4

Performance dimension 2 0 1 2 3 4

Performance dimension 3 0 1 2 3 4

etc

The following rater characteristics may be analysed:

a) A judge may be relatively harsh overall; estimates of this are available. Such estimates represent averages over all performance dimensions.

b) Judges may differ from each other in the way they interpret the steps on the scale. For example, one judge may use the full range of score points, another may use only two or three; these latter may be the middle of the range, or the extremes.

c) The point under (b) may be considered as a tendency in general, that is, across all performance dimensions, or may be considered separately for each performance dimension; that is, a judge may interpret the scale in one way for one performance dimension but differently for another.

Thus the analysis is capable of providing more or less general (conversely, more or less fine-grained or specific) statements about rater characteristics.

These general points will now be illustrated by an analysis of a real data set, to demonstrate the types and extent of variation among raters which can be modelled by the analysis. The research and practical implications of the analysis will then briefly be considered.

4 Methodology

a) The IELTS writing test

The International English Language Testing System (IELTS) (Alderson, 1988; Ingram, 1990; Griffin, 1990) is a joint British-Australian test of English for Academic Purposes, replacing the British Council ELTS test (Carroll, 1980) in late 1989. It is primarily intended as a screening test for university selection. The skills of listening and speaking are tested in general, non-academic

contexts, while modules assessing reading and writing are academically-oriented and subject-area specific. There are separate modules for (broadly) Arts and Social Sciences, Medical and Life Sciences, and Science and Technology; there is also a General Training module for students coming on non-academic, training attachments in an English speaking environment.

The writing test in the academic modules comprises two Tasks. Task One is an information transfer task. Candidates are required to produce a text of at least 100 words on the basis of a stimulus consisting of diagram or other graphic representation of information. Task Two is an essay-writing task, requiring the production of a text of at least 150 words in response to a stimulus consisting of one or more reading passages in the reading comprehension section of the test.

Performance on the tasks is rated in either of two ways, each involving an effective 9-point rating scale. Judges are normally given a choice of method. In the first method, scores are given on three performance dimensions for Task 1, four dimensions in Task 2. These are set out in Table 1.

Table 1 Performance dimensions, IELTS writing tasks

Task 1

Task fulfilment
Coherence and cohesion
Sentence structure

Task 2

Communicative quality
Arguments, ideas and evidence
Word choice, form and spelling
Sentence structure

Overall scores for each task are then obtained by averaging (and rounding up). In the second method, judges allocate a single overall score directly by consulting the global Band descriptors. Task 1 and Task 2 scores are then automatically converted to an overall Band score by means of a table which weights the tasks appropriately. In this study, judges were required to use the first method. In their normal practice most said that time constraints forced them to use the second method, even though they preferred the first, although this preference may not be generalizable to examiners as a whole.

Nine levels of performance, numbered from 1 to 9, are defined within each dimension. (The global Band scales

Figure 2 Band descriptors for performance dimensions, Task 1

[BLANK - of TEXT p 7]

are a simple composite of the relevant dimension band scales). The Band descriptors for these performance levels for Task 1 are shown in Figure 2. (These descriptors remain confidential and are not shown in publicly available versions of this paper)

b) Data

Performances on two IELTS written tasks from 49 candidates taking the Arts and Social Sciences module of the IELTS test at Melbourne test centres in November and December 1990 were marked independently by four raters. The raters were all accredited and experienced as raters of the IELTS writing test.

Counts were made of scores on each dimension for each judge. Inspection of the data showed that for Task 1, no ratings had been given in Bands 1, 2 and 9. Band 3 had been used by only one judge, on three occasions. For Task 2, no ratings had been given in Bands 1 or 9. Band 2 had been used by only one judge, on 9 occasions (involving 5 candidates); Band 3 had not been used by all judges for all dimensions. On the basis of these counts, it was decided to re-code the data in the following way:

Raw score (Band)	*Coded as*
2,3,4	0
5	1
6	2
7	3
8	4

thus reducing the scale used to a five-point scale. Data were thus available for performance on a five-point scale on three dimensions for Task 1 and on four dimensions for Task 2.

c) Types of analysis

A number of analyses were carried out using the FACETS programme. The programme allows one to choose between the use of the Rating Scale and the Partial Credit model for either the rater, or the performance dimensions, or both. This leads to a number of possible combinations, and therefore of types of analysis:

Rating Scale model for both performance dimensions and judges. This assumes that all judges across all performance dimensions are interpreting the thresholds on the rating scale in a uniform way.

Rating scale model for performance dimensions, Partial Credit model for judges. This assumes each judge is interpreting the rating scale in a uniform way across performance dimensions, but allows for variation between judges in their interpretation of the rating scale. Thus, one judge may use the middle scoring points (Bands 5 and 6) for the majority of candidates, regardless of

which performance dimension she is considering, whereas another may use the full set of score points, again consistently for all performance dimensions.

Rating scale model for judges, Partial Credit model for performance dimensions. This assumes that for a particular performance dimension, there is no variation between judges in their interpretation of the scale; but there may be variation from one performance dimension to the next. Thus, all judges may use the middle score points for let us say the dimension *Coherence and cohesion* but the full range of score points for the dimension *Sentence structure*, and so on.

Partial Credit model for both performance dimensions and judges. This makes the fewest assumptions: it allows for variation between judges in the way the scale is interpreted for any performance dimension, and also allows for variation across performance dimensions. Thus, one judge may concentrate on the middle score points for the performance dimension *Coherence and cohesion* but use the full range of score points for the dimension *Sentence structure*, while the exact reverse may be true for another judge.

In general, the difference between the models is a question of whether we ignore variation across performance dimensions or between judges (or not) when we consider the way the rating scale thresholds are being interpreted. Analysis of the same data set using the different models enables us to see how much variation is being disguised by this averaging out. We will also see that each of these analyses is valuable in providing different kinds of insight into the behaviour of raters.

5. *Analysis of Task 1*

We conducted a number of Rasch-based and non-Rasch-based analyses of data from Task 1. These analyses are seen as complementary, each capable of producing potentially useful and relevant information. First, a conventional inter-rater reliability analysis was done. The correlation matrix is reported in Table 2.

Table 2 ***Inter-rater reliability, Task 1 (Pearson)***

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*
Judge 1	1.000			
Judge 2	0.695	1.000		
Judge 3	0.566	0.731	1.000	
Judge 4	0.710	0.748	0.734	1.000

The analysis reveals a modestly respectable degree of inter-rater reliability on this measure; Judge 1 is identified as being somewhat less reliable than the other judges. Treating the dimensions as items, and using the uncoded bands with scores in the range 3-8, individual candidates could score a minimum total of 9, a maximum of 24. Table 3 reports descriptive statistics for the four judges on this basis.

Using the Partial Credit model for both judges and performance dimensions, 4 candidates were found to be misfitting, one on two performance dimensions. The mean ability of candidates was -0.53 logits, with a standard deviation of 2.16 logits. The index of reliability of person separation was .95. *Sentence structure* was the most harshly marked dimension, confirming the findings of McNamara (1990) and Pollitt and Hutchinson (1987), with a mean difficulty of 0.43 logits; less harshly marked was *Task fulfilment* (-0.06 logits), with *Coherence and cohesion* the most leniently marked dimension (-0.36

Table 3 Descriptive statistics, Judges, Task 1

	Judge 1	*Judge 2*	*Judge 3*	*Judge 4*
No. of cases	49	49	49	49
Mean (raw scores)	17.245	17.490	16.184	17.306
Standard deviation	3.591	2.123	2.224	3.117
Judge severity (logits)	-0.04	-1.01	1.03	0.01
Standard error	0.11	0.14	0.14	0.13

logits). The severity of judges is reported in Table 3. The results coincide with those from the analysis of raw scores reported in the same Table for the harshest and most lenient judges, but reverse the order of harshness for the middle two judges, who were fairly close together. Typically, calibrations of harshness produced by the model must be consistent with raw scores; the discrepancy in this case is due to the fact that the programme, as is usual in Rasch analyses, edits out data from candidates with perfect scores or scores of zero, so that the data on which the Rasch analysis is based differ somewhat from the data used in the conventional analyses.

Figure 3 shows ICCs resulting from an analysis in which the *Rating Scale* model was used for *performance dimensions*, and the *Partial Credit* model for *judges*. As stated above, this assumes that each individual judge will interpret the scale thresholds in the same way

Figure 3 Item characteristic curves, Rating Scale model used for performance dimensions, Partial Credit model used for judges, Task 1

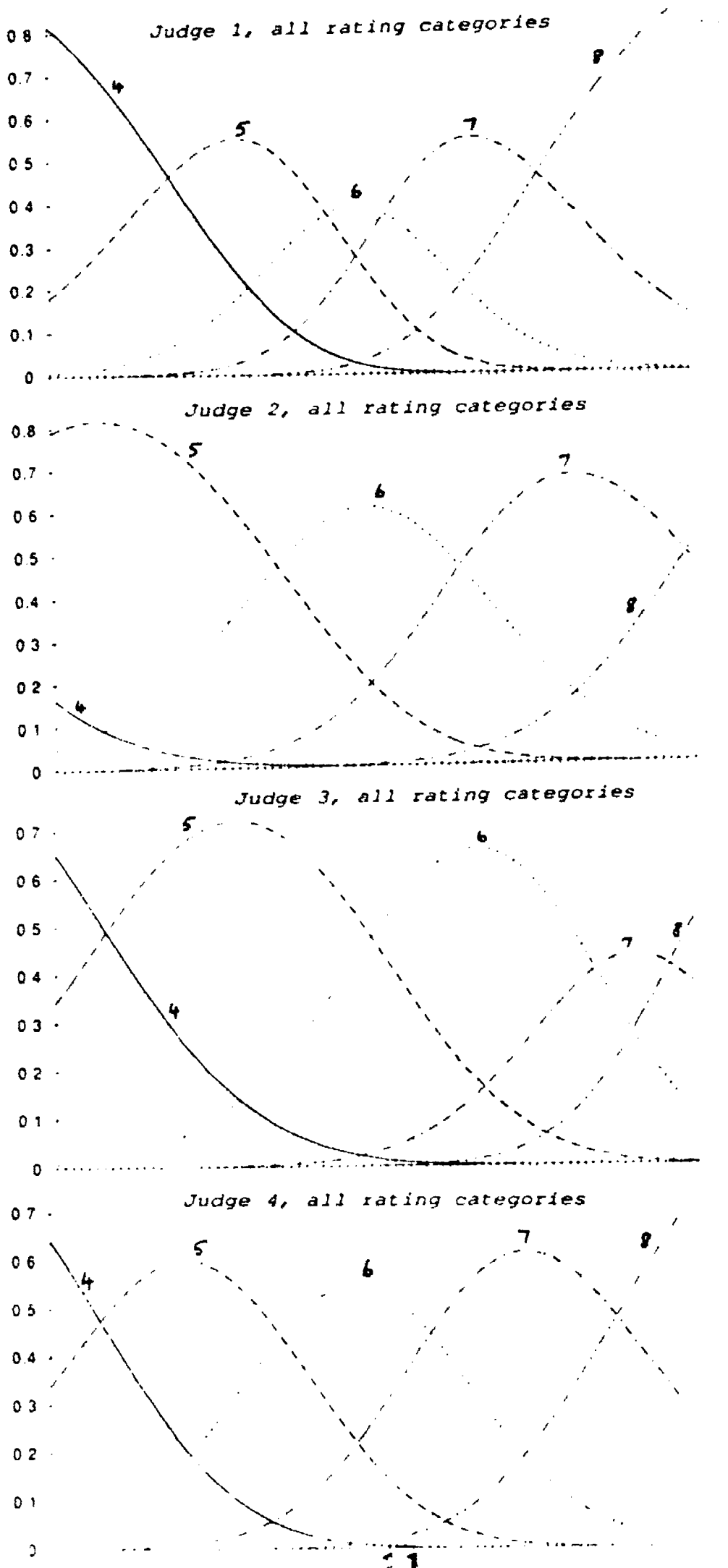
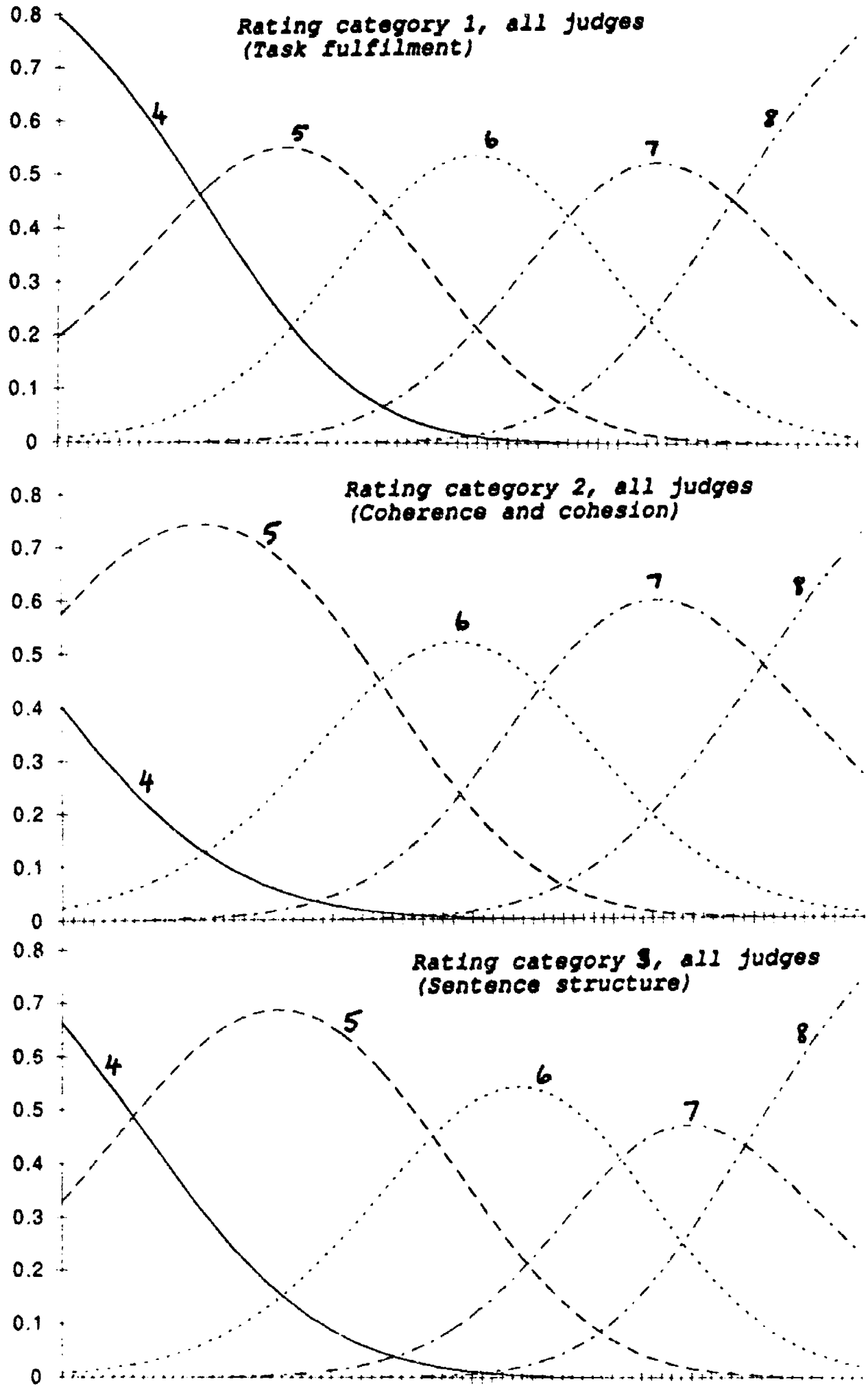


Figure 4 Item characteristic curves, Partial Credit model used for performance dimensions, Rating Scale model used for judges, Task 1



across all performance dimensions, and allows us to summarize the way judges differ from each other on average in their interpretation of the rating thresholds. Figure 3 shows interesting variation in the way in which judges interpret the rating scale in general. For example, Judges 1 and 4 use the whole width of the scale, whereas Judges 2 and 3 confine assessments largely to the central bands. Judges 1 and 4 contrast, however, in that Judge 1 is more prepared to use the extreme points of the scale (Bands 4 and 8) than Judge 4. In the crucial rating area around Band 6, candidates of a given ability are only slightly less likely to score a Band 5 or a Band 7 than a Band 6 with Judge 1. In general, Judges 1 and 4 discriminate between candidates more clearly than the other two judges. This helps us to define a possible type of rating style: 'definite' vs 'nuanced'.

The ICCs also illustrate the greater harshness of Judge 3 and the greater leniency of Judge 2 relative to the other judges. However, the leniency of Judge 2 is quite specific: it is mainly a question of that judge's use of Band 5 to cover a range of ability awarded either Band 4 or Band 5 by the other judges. In other words, the analysis allows us to be specify more exactly the nature of harshness or leniency in a judge.

Figure 4 shows ICCs for the analysis in which the *Rating Scale* model was used for *judges*, and the *Partial Credit* model for *performance dimensions*. It demonstrates again the relatively greater harshness of judges when rating on the dimension *Sentence structure* than when rating on other dimensions.

The analyses so far have provided information about rating styles, and may enable us to establish an inventory of rater types with definable rater characteristics. This could be valuable for research, and for rater training.

To what extent are such analyses obscuring variation associated with individual performance dimensions (Figure 3) or individual judges (Figure 4)? A *full Partial Credit analysis for both judges and performance dimensions* was carried out, and shows how each judge was using the score points for each performance dimension. The analysis revealed an interaction between judge and performance dimension in the interpretation of thresholds on the rating scale. That is, there were considerable differences in the way score point thresholds were being interpreted both between judges for particular performance dimensions and between performance dimensions for particular judges. The 12 associated ICCs (4 judges x 3 performance dimensions) are shown in Figure 5.

Looking first at graphs across the columns, that is, row-wise, we can examine the way in which individual judges vary their interpretation of scale thresholds across performance dimensions. Judge 4, for example, provides

Figure 5 Item characteristic curves for full Partial Credit analysis for both judges and performance dimensions, Task 1

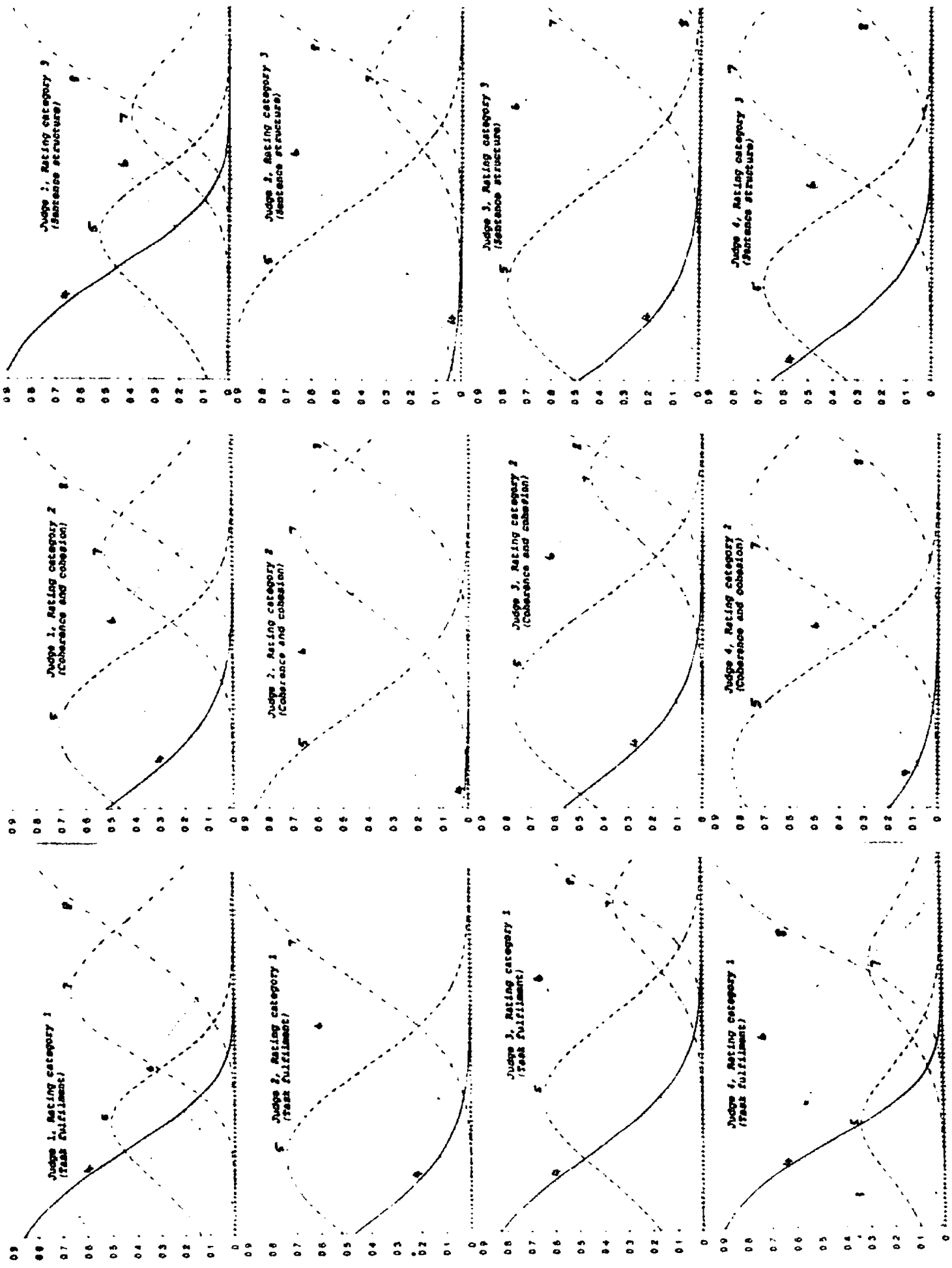
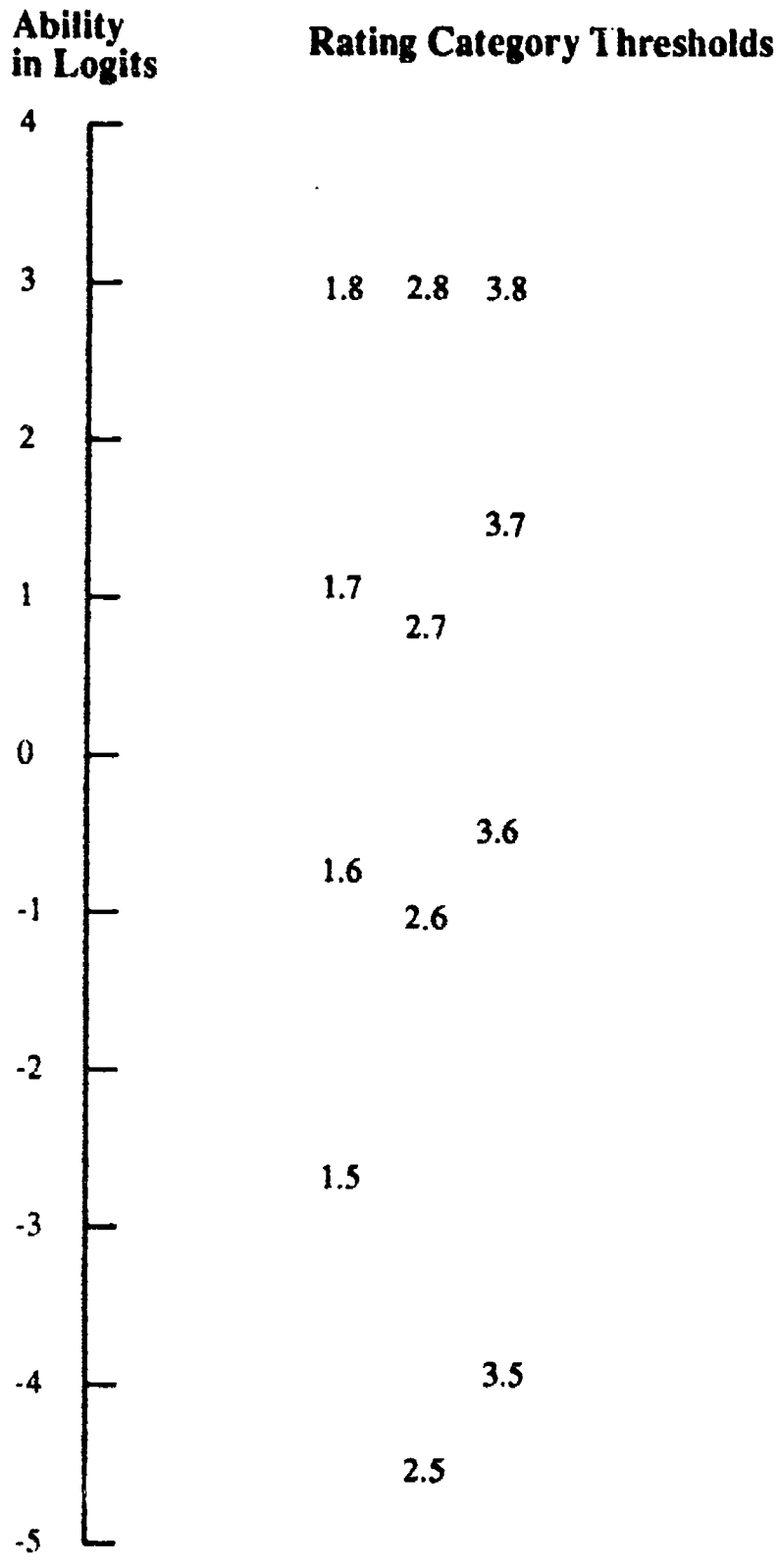


Figure 6 Rating category thresholds, Task 1



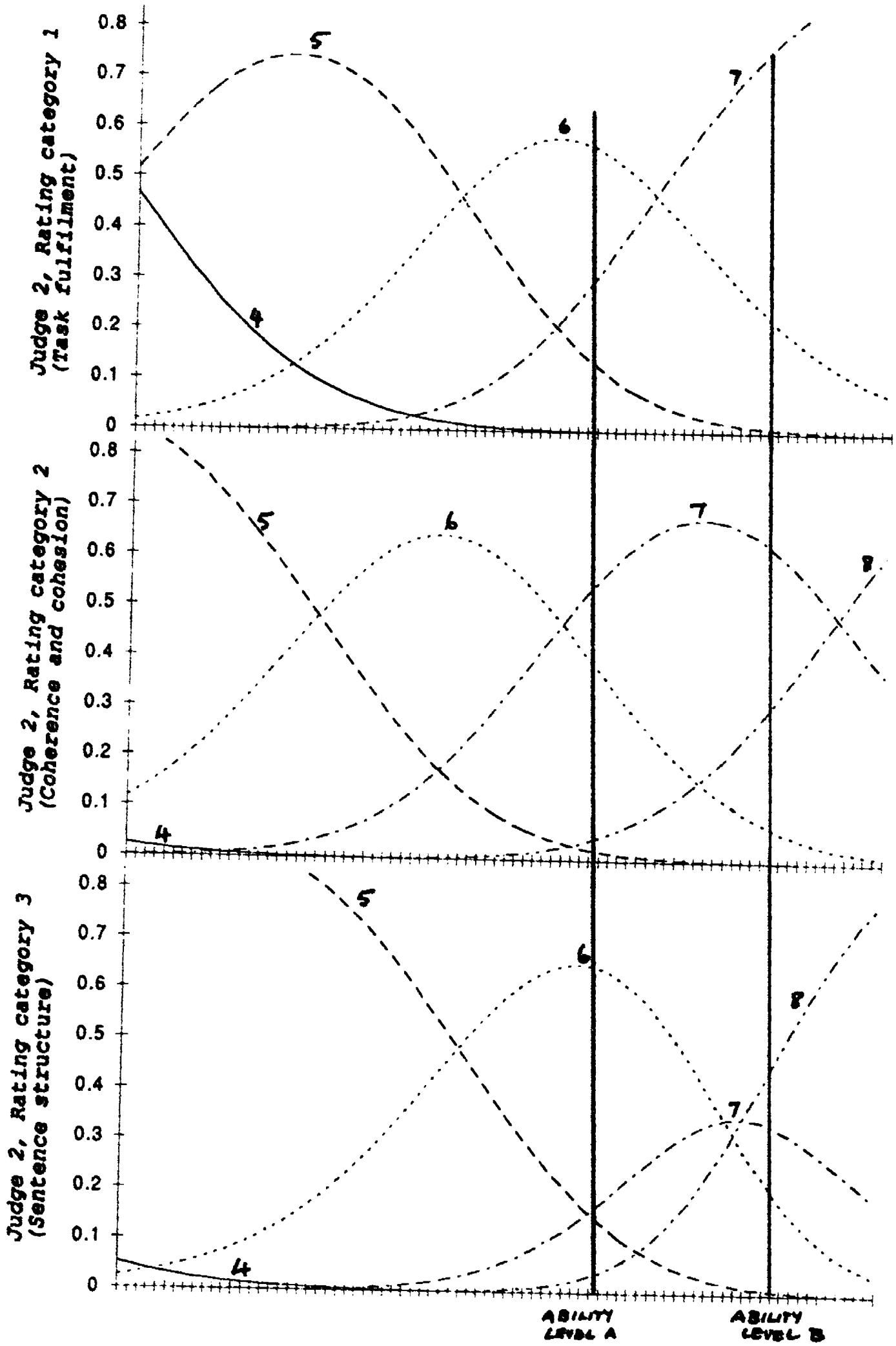
an interesting contrast in her interpretation of the scale thresholds between Performance dimension 1 (*Task fulfilment*) on the one hand and Performance dimensions 2 and 3 (*Coherence and cohesion* and *Sentence structure*) on the other. For Performance dimension 1, a candidate in the ability range where the most likely score is a Band 5 is almost as likely to get a Band 4 or a Band 6; with another performance dimension, however - Performance dimension 2 - such a candidate would certainly get a 5. For Performance dimension 1, the greatest likelihood of getting a score of Band 7 is associated with an ability range in which scores of Band 6 or Band 8 are equally (actually, more) likely; candidates of the same ability would almost certainly score a 7 for Performance dimensions 2 and 3. For Judge 2, her interpretation of the scale at Bands 7 and 8 differs markedly from Performance dimension 1 to Performance dimension 3.

The analysis also permits us to throw some light on the issue of the componential structure of ability at a given level. It has often been pointed out that an overall Band score of, say, 7 may represent unevenness in the separate abilities contributing to the aggregate score. For example, a person may be relatively weak in one area, relatively strong in another, and these balance each other out; in another candidate these things may be reversed, but with the same effect of balancing out. Figure 6 represents data from the analysis in which variation between judges has been ignored. It shows the ability level at which a candidate has a 50% chance of having mastered the skills required to score a particular Band (or above) for a particular performance dimension. Thus, a candidate of ability +1 logit has a 50% chance of scoring Band 7 or above on Performance dimension 1. A candidate of ability +3 logits has a 50% chance of scoring an 8 on each of the three performance dimensions.

It will be observed that the performance dimensions roughly cluster together at ability levels equivalent to Bands 6, 7 and 8. However, at Band 5 the picture is less uniform. A person who at ability level -4 logits has a 50% chance of scoring Band 5 on each of Performance dimensions 2 and 3, but is most likely to score Band 4 on Performance dimension 1. To the extent that Band 5 represents a likely cut-off score, failure to provide a profile of the student's ability may disguise significant variation across performance dimensions. But on the whole, ignoring the effect of particular judges, the performance dimensions do seem to cluster at Band levels, suggesting that the grouping together of performance dimensions into a single statement of ability makes empirical sense.

If we consider again part of the data from analysis which treats each judge separately, this picture begins to break down somewhat. Figure 7 represents an analysis of the behaviour of Judge 2. The vertical line marked Ability Level A represents a given ability level. The

Figure 7 ICCs from Judge 2, Task 1



most likely profile generating this particular ability measurement is a Band 6 on *Task fulfilment*, a Band 7 on *Coherence and cohesion* and a Band 6 on *Sentence structure*. Now look at the line marked Ability Level B, indicating a higher level of ability. Here the profile has changed, so that for the same judge the most likely contribution to the definition of this ability will be Band 7 on *Task fulfilment*, Band 7 on *Coherence and cohesion* and Band 8 on *Sentence structure*. (Of course this analysis represents a generalization, too, this time about all candidates for this particular judge). The analysis presented here may be seen as an argument in favour of profile reporting, despite its relative unpopularity, if we think that wording of the band descriptors is going to be taken seriously. This profile of ability structure may vary across raters, moreover.

Returning to Figure 5, if we look *down* the columns, we see the variation across judges for particular performance dimensions. The variation across judges which we noted overall in the data presented in Figure 4 is now compounded by the fact that this variation will differ for particular performance dimensions.

We have thus established the fact of variation across judges. But how significant is this variation? Two procedures were devised for examining this issue.

First, ability estimates for candidates were derived for each judge at a time, without any further marking. A correlation matrix of ability estimates from the four judges was derived (Table 4). (Disattenuated estimates are reported in Table 5).

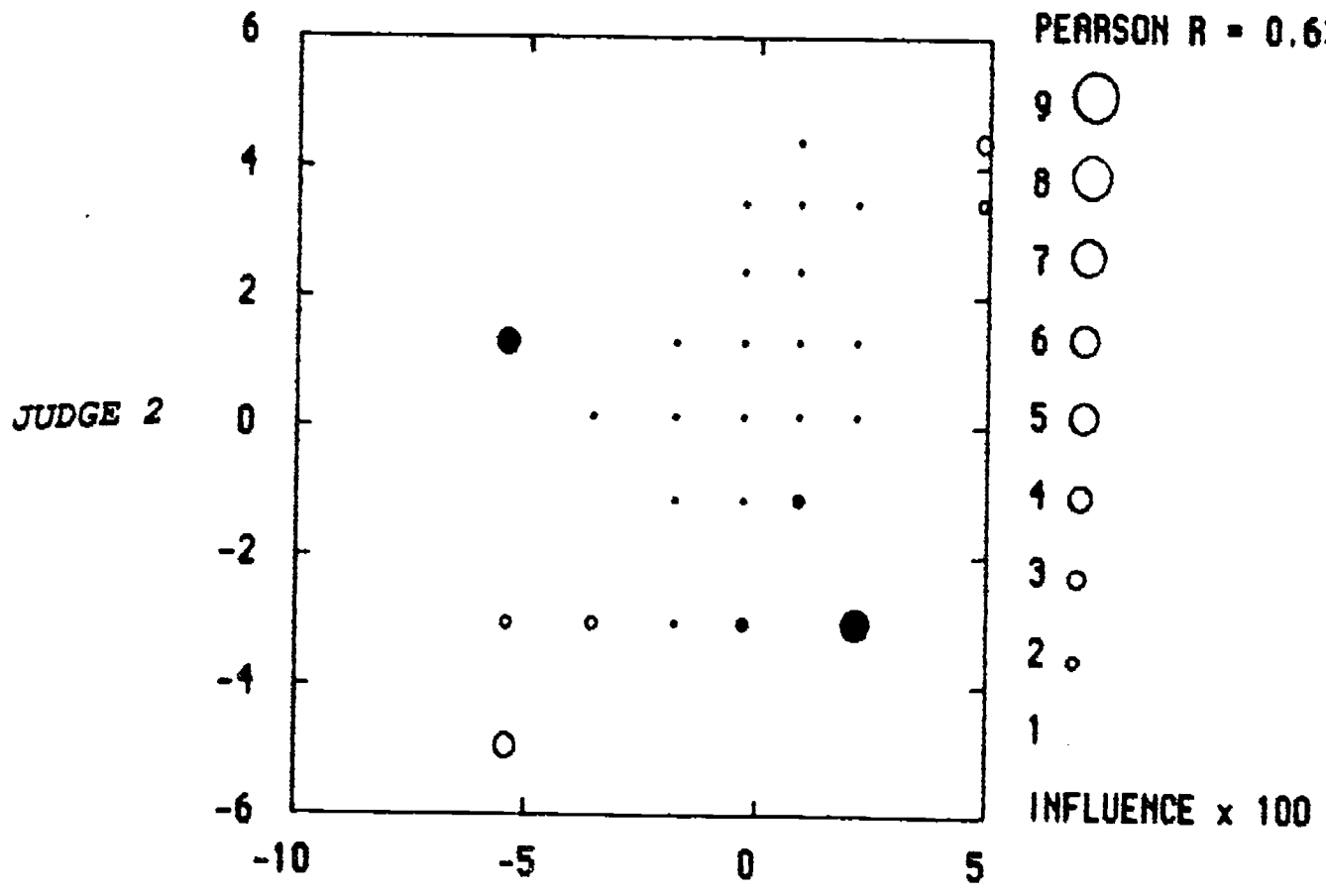
Table 4 Correlation coefficients (Pearson) for pairs of ability estimates, 4 judges

	Judge 1	*Judge 2*	*Judge 3*
Judge 2	.586 *n=45*		
Judge 3	.419 *n=43*	.630 *n=45*	
Judge 4	.373 *n=36*	.585 *n=38*	.585 *n=37*

Table 5 Disattenuated correlation coefficients (Pearson) for pairs of ability estimates, 4 judges

	Judge 1	Judge 2	Judge 3
Judge 2	.666 n=45		
Judge 3	.505 n=43	.808 n=45	
Judge 4	.401 n=36	.672 n=38	.705 n=37

Figure 8 Plot of ability estimates from two judges working alone, Task 1



The sets of ability estimates were plotted against each other to identify outliers, as in Figure 8. Such outliers would be candidates at serious risk of mismeasurement from one judge working alone.

Second, the estimates derived from data from a single judge were compared with the best available estimates of ability based on data from all four judges. A *chi-*squared test was developed to test for the significance of these deviations. This test proved to be massively significant. The reason for this was that the actual values of measurement of ability showed the influence of the extent to which particular judges discriminated between candidates. For example, Judges 1 and 4 discriminated between candidates more than Judges 2 and 3, and this was reflected in a greater spread of ability estimates for candidates on the logit scale.

6 Analysis of Task 2

There was some inconsistency in the way raters handled texts in Task 2 which were deemed to be short. First, interpretation of what *constituted* a short text varied among raters; for example, one rater identified 11 scripts as being short, another identified only one. Secondly, raters varied in their *treatment* of those texts identified as being short. Some raters marked the candidate using the criteria in the normal way, as if the text was a full length text, converted the marks to an overall band score, then reduced the candidate's overall score by one band. Other raters marked the candidate down by one band in the dimension *Arguments, ideas and evidence* only. For comparability of data, and for simplicity, it was decided to proceed as if this latter procedure had been adopted by all raters, and the scores of the raters adopting the former procedure were amended accordingly.

The main issue addressed in the analysis of data from Task 2 was the extent to which raters behaved consistently or differently across tasks. An analysis of Task 2 was carried out using the Partial Credit model for both judges and performance dimensions. This time only 2 candidates were found to be misfitting. The mean ability of candidates was -0.87 logits, with a standard deviation of 2.97 logits. The index of reliability of person separation was .97. Compared with Task 1, candidates found this task harder, and there was greater variation between candidates in terms of performance.

Table 6 shows the difficulty estimates for dimensions. *Word choice* was the most leniently scored dimension (-0.79 logits); *Communicative quality* and *Sentence structure* were roughly equivalent in difficulty (-0.04 and -0.01 logits respectively) while the most difficult dimension was *Arguments, ideas and evidence*. It is not clear what effect the penalizing for shortness had on the

estimate for the difficulty of this dimension; but it is worth noting that there was significant misfit for this dimension.

Table 6 Difficulty estimates, performance dimensions, Task 2

Performance dimension	*Difficulty (logits)*	*Error*	*Infit*	
			MnSq	*Std*
Communicative quality	-0.04	0.13	0.9	0
Arguments, ideas and evidence	0.84	0.13	1.3	2
Word choice	-0.79	0.13	0.9	0
Sentence structure	-0.01	0.13	0.9	0

Interestingly, the overall severity of the judges was different for this task, revealing a task-judge interaction. The severity estimates for judges on each task are reported in Table 7.

Table 7 Estimates for judges, Tasks 1 and 2

Judge	*Task 1*	*Rank*	*Task 2*	*Rank*
1	-0.04	3	0.57	1
2	-1.01	4	-0.32	3
3	1.03	1	0.25	2
4	0.01	2	-0.50	4

A statistical test of these discrepancies proved significant in each case. In other words, the analysis revealed clear differences in rater behaviour across the two tasks.

7 *The effect of re-marking on ability estimates*

The issue of re-marking

At present, IELTS scripts are rated by a single trained and accredited rater; double rating is not carried out, presumably on the grounds of expense. 10% of scripts from any one rater, sampled at random, are in fact examined again by another rater as part of a process of monitoring the standards of raters, but this may be done some time after a particular session of the test, and will not affect candidates' scores directly. The question of the advantages of double marking have of

course long been discussed; for a recent discussion from the point of view of generalizability theory, see Bachman (1990), who also summarizes the findings of van Weeren and Theunissen (1987). The latter study considered the question of the effect of more than one rating of performances on a pronunciation task of *all* candidates. In the present study, the effect of the re-marking of only a *proportion* of scripts by a second rater is considered.

It is possible to investigate this issue using multi-faceted Rasch analysis because of the way the calibration of the ability of individual candidates is done. We may understand this by comparing Rasch ability estimates from FACETS with raw score estimates. Using raw scores only, it would not make sense to re-mark only a proportion of scripts selected *at random* (that is, not only difficult or borderline cases) as the benefit of double marking would be restricted to the lucky few who were chosen; they would not affect the marks of candidates whose scripts were marked by a single rater. The ability estimates derived from FACETS, however, are made on the basis of information about an individual's performance on a set of items *and* information about the raters themselves. In other words, the whole data matrix is involved in the calibration of the ability of individuals. We may conceptualize the influence of double marking as one of progressively enriching the data matrix by providing progressively more information about the characteristics of raters, which in turn will affect the estimates of all candidates.

Methodology

The scores of two raters on Task 1 were selected for the study. It was decide to use Judges 2 and 3. Judge 1 was excluded because scores from this judge had been found to be less reliable than the others (cf Tables 2 and 5). Judge 4 was excluded as this judge used 'extreme' scores (maximum and minimum possible) several times, and such scores are routinely excluded in Rasch analyses (cf figures for number of candidates included in each correlation in Tables 4 and 5). It was noted that the two judges chosen (Judges 2 and 3) differed more than any other pair of judges in terms of harshness (cf Table 7).

A programme was written which took the data from the two chosen judges and sampled from it at random to include various proportions of double marking. FACETS analyses were then run on the successive data sets. The following proportions of double marking were used: 0%; 10%; 20%; 30%; 40%; 50%; 60%; 70%; 80%; 90%. The sampling procedure was repeated, and a set of FACETS runs again carried out. An average was then taken of the ability estimates for individual candidates from the two sets of runs. These average ability estimates from the successive partial data sets were then compared with those based on the complete data set (that is, using

information from all four judges). These latter estimates represented 'best' estimates of candidates' abilities.

The ability estimates from the partial data sets ('partial estimates') and the 'best' estimates were then compared in two ways.

First, the correlations between the sets of 'partial' and 'best' estimates were linearized using Fisher's z transformation, and plotted against each other. The linearization was done in order to demonstrate more clearly any changes in correlation resulting from changes in the proportion of double marking). Second, the actual values of the ability estimates were compared and the root mean square error of the 'partial' ability estimates was plotted against the proportion of double-marking.

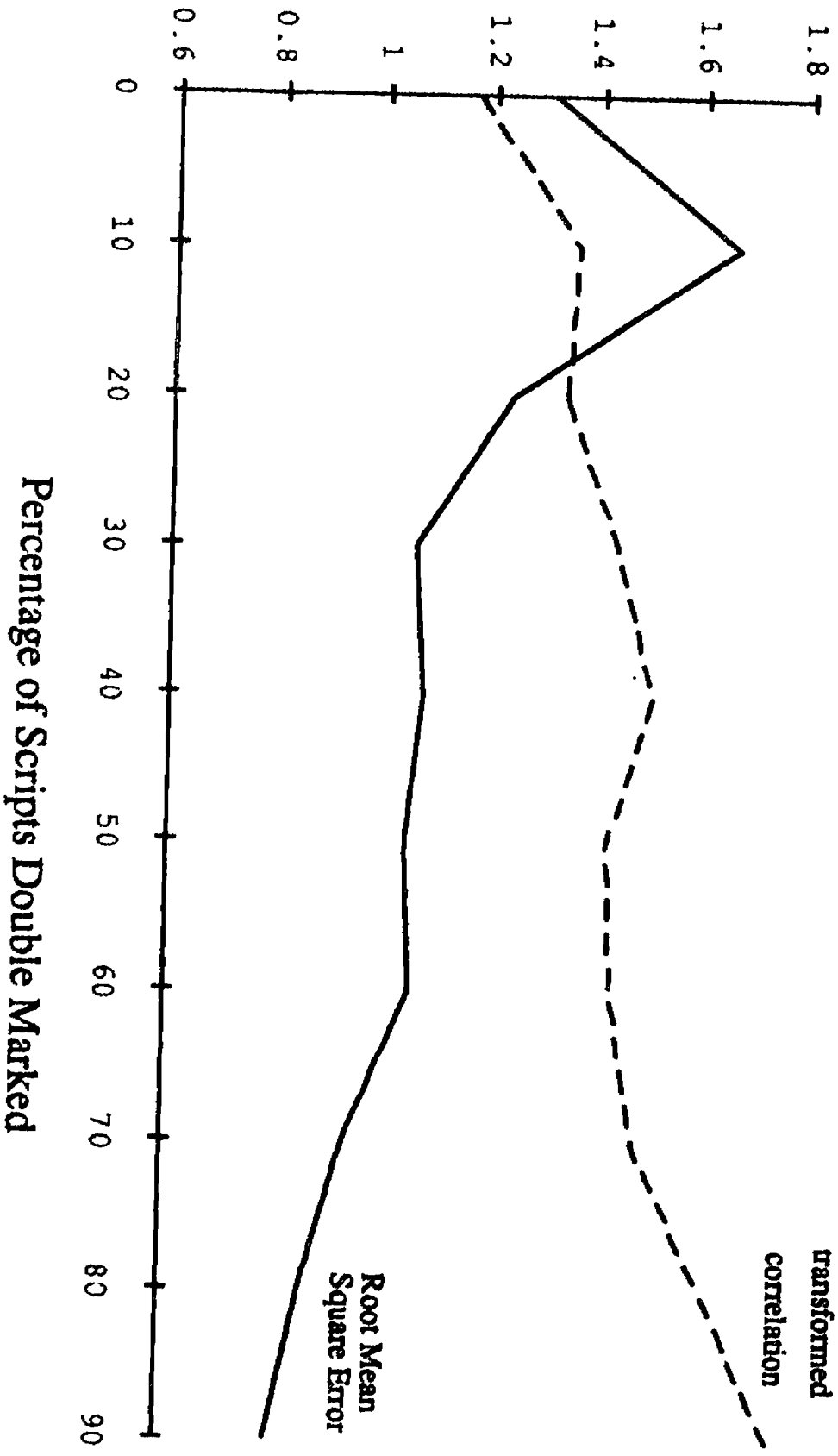
Results

The results of both analyses are shown in Figure 9.

It had been hypothesized that the benefit to be gained from double marking might as it were flatten out after a certain proportion of scripts had been double marked; that is, there might be a point of diminishing returns. This would represent a potential saving, for example if re-marking let us say 50% of scripts did not provide any more information than re-marking 90% of the scripts. The point of maximum fluctuation in the stability of ability estimates was also of interest.

In fact, an approximately linear relationship was discovered. (It is probable that a larger study would confirm that the observed relationship was indeed linear. The amount of random sampling done in this study was very limited; in a true Monte Carlo study hundreds or thousands of random samples would be taken - in this case only two sets of random samples were used). That is, the more double marking was done, the better the ability estimates were, and this applied across the whole set of proportions of double marking. There is no law of diminishing returns, it seems. Nevertheless, Figure 9 reveals that estimates will improve in direct proportion to cost; the more money you spend, the better the estimates will be. This is an unsurprising finding, but nevertheless perhaps acts as a salutary reminder that reliability does not come cheaply. Conversely, however, assuming that the relationships reported are in fact linear, then even with 10% double marking, we are taking rater effects into account and getting better estimates than we would using raw scores or deriving estimates from the marking of a single rater only.

Figure 9 The effect on ability measures of varying proportions of double marking of scripts



8 Further research

The FACETS programme allows us to estimate rater harshness, and the estimates of ability compensate for this. An unresolved question is whether these estimates will remain constant for raters over time. We have already seen that raters vary in their harshness from Task to Task; this therefore suggests that there may indeed be variability over time; this needs to be investigated. If rater harshness were found to be constant, then compensation for this could be built into ratings from particular judges at future test sessions. Practically speaking, calibrations of rater harshness could be derived from the training stage, and then incorporated into ratings from particular judges as they proceeded to work. Alternatively, feedback could be given to raters in training as to the characteristics of their rating style, in order that where appropriate this style might be modified.

A further area of research would be an extension of the study of double marking into a true Monte Carlo study, to confirm the finding of linearity reported above. More work needs to be done, too, on the componential structure of overall Band scores, along the lines suggested in Section 5 above.

Finally, much more work needs to be done on the definition of consistent rater types. In this study, we raise the possibility that raters may be characterized as, for example, 'definite' vs 'nuanced'. Many more raters need to be studied in order to confirm that such characterizations (and others) are meaningful.

9 Conclusion

The present study has demonstrated the usefulness of multi-faceted Rasch measurement as a research tool for language testers, and has suggested the possibility of its practical applications. An analysis of real data from the IELTS writing test has demonstrated the way in which consistency and variability in rater behaviour may be revealed and explored, at varying levels of specificity and generality. This has implications for research on rater types, and practical implications for rater training; and raises the possibility of adjusting marks to take account of rater effects. The study has also demonstrated how the new Rasc techniques may be used to address questions of economy and reliability in the double marking of candidates' scripts. The potential for further theoretical and practical language testing research using this new technique is considerable.

NOTE

[1] This is a revised version of a paper originally presented at the 13th Language Testing Research Colloquium, Educational Testing Service, Princeton, NJ, 21-23 March 1991.

REFERENCES

Alderson, J.C. (1988). New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing* 5,2: 220-232.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika* 43: 561-573.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement* 38: 665-680.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Carroll, B.J. (1980). *Testing communicative performance.* Oxford: Pergamon.

Griffin, P.E. (1990). *The statistical properties of the IELTS test battery - Australian data.* Paper given at the RELC Seminar on 'Language Testing and Language Programme Evaluation', Singapore, 9-12 April.

Ingram, D.E. (1990). *The International English Language Testing System (IELTS): its nature and development.* Paper presented at the RELC Seminar on 'Language Testing and Language Programme Evaluation', Singapore, April 9-12.

Krzanowski, W.J. and A.J. Woods (1984) Statistical aspects of reliability in language testing. *Language Testing* 1,1: 1-20.

Linacre, J.M. (1989). *Many-faceted Rasch measurement.* Ph.D. thesis, University of Chicago.

Linacre, J.M. (1990). *FACETS computer program for many-faceted Rasch measurement (Version 2.36).* Chicago, IL: MESA Press.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47: 149-174.

Masters, G.N. and B.D. Wright (1984) The essential process in a family of measurement models. *Psychometrika* 49: 529-544.

McNamara, T.F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing* 7,1: 52-76.

Pollitt, A. and C. Hutchinson (1987). Calibrated graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing* 4,1: 72-92.

Van Weeren, J. and T.J.J. Theunissen (1987). Testing pronunciation: an application of generalizability theory. *Language Learning* 37,1: 109-122.

Woods, A., P. Fletcher and A. Hughes (1986) *Statistics in language studies.* Cambridge: Cambridge University Press.

Wright, B.D. and G.N. Masters (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B.D. and M.H. Stone (1979). *Best test design.* Chicago: MESA Press.